REPORT DOCUMENTATION PAGE

AFRL-SR-AR-TR-02-

0357

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
16 Oct 02

3. REPORT TYPE AND DATES COVERED
FINAL REPORT 15 DEC 00 TO 14 DEC 02

4. TITLE AND SUBTITLE
[redacted]

5. FUNDING NUMBERS
F49620-01-1-0098

2312/AX

61102F

6. AUTHOR(S)
DR SUBHASH C. BASAK

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITY OF MINNESOTA
CENTER FOR WATER AND THE ENVIRONMENT
5013 MILLER TRUNK HIGHWAY
DULUTH MN 55811

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NL
4015 WILSON BLVD., ROOM 713
ARLINGTON, VA 22203-1954

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
APPROVE FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

In recent years, there has been increased interest in the development and use of quantitative structure activity/property relationship (QSAR/QSPR) models. For the most part, this is due to the fact that experimental data is sparse and obtaining such data is costly, while theoretical structural descriptors can be obtained quickly and inexpensively. In this study, three linear regression methods, viz, principal component regression (PCR), partial least squares (PLS), and ridge regression (RR). were used to develop QSPR models for the estimation of human blood; air partition coefficient (logP blood;air) for a group o: 31 diverse low-molecular weight volatile chemicals from their computed molecular descriptors. In general, RR was found to be superior to PCR or PLS. Comparisons were made between models developed using parameters based solely on molecular structure and linear regression (LR) models developed using experimental properties, including saline;air partition coefficient (longP saline;air) and olive oil;air partition coefficient (logP olive oil;air), as independent variables, indicating that the structure-property correlations are comparable to the property-property correlations. The best models, however, were those which used rat logP blooda;air as the independent variable. Haloalkane subgroups were modeled separately for comparative purposes, and although models based on the congeneric compounds were superior, the models developed on the complete set of diverse compounds were of acceptable quality. The structural descriptors were superior, the models developed on the complete set of diverse compounds were of acceptable quality.

14. SUBJECT TERMS
Blood:air partition coefficient; PBPK model; theoretical molecular descriptors; ridge regression; quantitative structure-property relationship (QSPR) model.

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclass	Unclass	Unclass	

1122 27

Submitted to *Risk Analysis*

PREDICTION OF HUMAN BLOOD:AIR PARTITION COEFFICIENT: A COMPARISON OF STRUCTURE-BASED AND PROPERTY-BASED METHODS

S. C. BASAK,[a] D. MILLS,[a] D. M. HAWKINS,[b] and H. A. EL-MASRI [c]

[a] *Natural Resources Research Institute, University of Minnesota Duluth*
5013 Miller Trunk Highway, Duluth, MN 55811, USA
[b] *School of Statistics, 313 Ford Hall, 224 Church Street S. E., University of Minnesota Minneapolis, MN 55455, USA*
[c] *Computational Toxicology Laboratory, Division of Toxicology Agency for Toxic Substances and Disease Registry (ATSDR), Executive Park Building 4, 1600 Clifton Road, E-29, Atlanta, GA 30333, USA*

In recent years, there has been increased interest in the development and use of quantitative structure activity/property relationship (QSAR/QSPR) models. For the most part, this is due to the fact that experimental data is sparse and obtaining such data is costly, while theoretical structural descriptors can be obtained quickly and inexpensively. In this study, three linear regression methods, *viz.* principal component regression (PCR), partial least squares (PLS), and ridge regression (RR), were used to develop QSPR models for the estimation of human blood:air partition coefficient ($\log P_{blood:air}$) for a group of 31 diverse low-molecular weight volatile chemicals from their computed molecular descriptors. In general, RR was found to be superior to PCR or PLS. Comparisons were made between models developed using parameters based solely on molecular structure and linear regression (LR) models developed using experimental properties, including saline:air partition coefficient ($\log P_{saline:air}$) and olive oil:air partition coefficient ($\log P_{olive\ oil:air}$), as independent variables, indicating that the structure-property correlations are comparable to the property-property correlations. The best models, however, were those which used rat $\log P_{blood:air}$ as the independent variable. Haloalkane subgroups were modeled separately for comparative purposes, and although models based on the congeneric compounds were superior, the models developed on the complete set of diverse compounds were of acceptable quality. The structural descriptors were placed into one of three classes based on level of complexity: Topostructural (TS), topochemical (TC), or 3-dimensional / geometrical (3D). Modeling was performed using the structural descriptor classes both in a hierarchical fashion and separately. The results indicate that the highest quality structure-based models, in terms of descriptor classes, were those derived using TC or TS+TC descriptors.

Key Words: Blood:air partition coefficient; PBPK model; theoretical molecular descriptors; ridge regression; quantitative structure-property relationship (QSPR) model.

1. INTRODUCTION

Modern lifestyle worldwide is based on the use of a large number of chemicals. Natural and synthetic chemicals are used as drugs, pesticides, herbicides, components of diagnostic tools, ingredients and solvents in industrial processes, to name just a few. The Toxic Substances Control Act (TSCA) Inventory maintained by the United States Environmental Protection Agency (USEPA) currently has over 81,000 entries and the list is growing every year.[(1)] Many of these chemicals are used for various purposes and have the potential to be released in the environment. Therefore, it is natural that we need to carry out risk assessment of the TSCA chemicals, particularly for those that are used frequently and in large quantities. Volatile organic chemicals (VOCs) constitute a class of chemicals that are frequently used in various industrial processes. Therefore, there is an interest to predict the potential adverse effects of these chemicals on human and environmental health. The overall risk of a chemical is determined primarily by its intrinsic toxicity (hazard) and exposure potential.

The blood:air partition coefficient of VOCs is an important determinant of pulmonary uptake of such chemicals from inhaled air. Such parameters are routinely used in building physiologically-based pharmacokinetic (PBPK) models for exposure assessment of such chemicals. Solubility of VOCs in blood is determined by its composition including the content of neutral lipid, phospholipid, and water, as well as the extent of binding of these chemicals to specific components such as plasma proteins and hemoglobin.[(2)] Such physicochemical considerations can be used to come up with physicochemically-based methods for the estimation of partition coefficient values of chemicals. The other possibility is the use of molecular descriptors to estimate partition coefficient of chemicals directly from their structure. Such quantitative structure-activity/property relationship (QSAR/QSPR) methods derived using theoretical descriptors are based on the idea that observable physicochemical and biological properties of chemicals are determined by their molecular structure. In particular, QSPRs have been found to be useful in the estimation of physicochemical properties such as octanol:water partition coefficient of various groups of chemicals,[(3, 4)] as well as the degree of transport through the blood-brain barrier[(5)] and skin,[(6)] of various congeneric and diverse sets.

While some quantitative models use experimental data per se as independent variables, it is important to note that experimental data does not exist for the majority of compounds, and obtaining such data is costly in terms of time and monetary resources. Computational modeling involving algorithmically calculated parameters based solely on molecular structure is an inexpensive alternative. In this paper, we have attempted to develop QSPR models to estimate human blood:air partition coefficients for a set of 31 VOCs using molecular descriptors which can be computed directly from molecular structure.

2. METHODS

2.1 Database. Liquid:air partition coefficients were experimentally determined by Gargas *et al.*[7] using a modified version of the gas-phase vial equilibrium technique[8] for a set of low molecular-weight volatile chemicals. Table I includes experimentally determined human and male Fischer 344 rat blood:air partition coefficient data for a set of 31 chemicals including 18 haloalkanes, 2 nitroalkanes, 2 aliphatic hydrocarbons, 4 haloalkenes, and 5 aromatics compounds. The human blood:air partition coefficient values were determined on blood pretreated with diethyl maleate to inhibit an observed glutathione transferase reaction. Experimental saline:air and olive oil:air partition coefficients, determined by Gargas *et al.*, are also listed in Table I. All experimental values were obtained at 37 °C.

It should be noted that the data used in the current study are a subset of that reported by Gargas et al.[7] Two cis/trans isomers were eliminated because they are indistinguishable in terms of their calculated molecular descriptors based on SMILES input. Methyl chloride was also removed from the data set as it is not possible to calculate our entire set of theoretic descriptors on two-atom compounds. In addition, two compounds were reported without discrete values for 0.9% saline:air partition coefficient and thus were not included in this study.

2.2 Theoretical Molecular Descriptors. Theoretical molecular descriptors may be divided into hierarchical classes based upon level of complexity. Topostructural (TS) descriptors, which encode information strictly on the adjacency and connectedness of atoms within a molecule, make up the simplest of the hierarchical classes. Topochemical (TC) descriptors encode information related to the chemical nature of a molecule including bond type. The 3-dimensional or shape descriptors (3D) are still more complex, encoding information about the 3-dimensional aspects of a molecule. Calculated $\log P_{n\text{-octanol:water}}$ descriptors[9] were included at the final stage of hierarchical model development. The topostructural and topochemical descriptors are collectively referred to as topological descriptors.

Descriptors used in the present study were derived from molecular structure using software packages including POLLY,[10] Triplet,[11, 12] and Molconn-Z.[13] From POLLY, a set of topological descriptors is available, including a large group of connectivity indices,[14-17] path-length descriptors,[14] and information theoretic[18, 19] and neighborhood complexity indices.[19] The Triplet descriptors also constitute a large group of topological parameters. They are derived from a matrix, a main diagonal column vector, and a free term column vector, converting the matrix into a system of linear equations whose solutions are the local vertex invariants. These local vertex invariants are then used in the following mathematical operations in order to obtain the triplet descriptors:

1. Summation, $E_i x_i$
2. Summation of squares, $E_i x_i^2$
3. Summation of square roots, $E_i x_i^{1/2}$
4. Sum of inverse square root of cross-product over edges ij, $E_{ij}(x_i x_j)^{-1/2}$
5. Product, $N(E_i x_i)^{1/N}$

Molconn-Z provides additional topological descriptors, including an extended set of connectivity indices, electrotopological indices,(20, 21) and hydrogen bonding descriptors, as well as a small set of molecular shape descriptors.

H-Bond, a software program developed by Basak,(22) was used to calculate HB_1, a measure of hydrogen bonding potential. Balaban's J indices were also calculated by software developed by the authors.(23-25)

$\text{LogP}_{n\text{-octanol:water}}$ values were calculated by the LogP program(9) and are included in Table I. Table II provides a brief description of all other theoretical molecular descriptors used in the current study, though the calculated values for these descriptors are not included for the sake of brevity.

2.3 Statistical Analysis. Independent and dependent variables were scaled by the natural logarithm, as their respective ranges differed by several orders of magnitude. The CORR procedure of the SAS statistical package(26) was used to identify perfectly correlated descriptors, i.e. $r = 1.0$. In each case, only one descriptor of a perfectly correlated pair was retained for use in the subsequent analysis. Any descriptor that either had a value of zero for all compounds in the data set or could not be calculated for all compounds in the data set was removed.

The structure-property models were developed using ridge regression (RR),(27) principal components regression (PCR),(28) and partial least squares (PLS) regression(29-31) methodologies, utilizing molecular descriptors in a hierarchical fashion. In addition, each class of descriptors was used independently to obtain single-class models. RR, PCR, and PLS are useful in cases wherein the number of descriptors is much greater than the number of observations, as well as in cases where the independent variables are highly intercorrelated. In addition, these regression methods make use of all independent variables as opposed to subset regression wherein it is possible that important parameters may be eliminated from the study. Linear regression (LR) was used to obtain the property-property models, which involve 1-2 independent variables. Statistical parameters reported include the cross-validated R^2 value and the PRESS statistic which are reliable measures of model predictability. In addition, the *t* values can be examined in order to identify significant descriptors. Although a descriptor with a large | *t* | indicates that the associated descriptor is important in the model, it should be cautioned that the reverse is not necessarily true.

Honest assessment of the quality of a prediction model is seldom straightforward, but is particularly challenging in a situation such as this where the number of independent variables far exceeds the number of observations.(32, 33) In these cases, conventional regression measures such as R^2 are useless. The measure we use is the cross-validation (or jack-knife) sum of squares. For this measure, each compound in turn is omitted from the data set, and the coefficients of the regression model (RR, PLS or PCR) computed using the remaining n-1 cases. These coefficients are used to predict the hold-out case. The overall quality of the fit is measured by the prediction sum of squares PRESS – the sum of squares of the difference between the actual observed activity and that predicted from the regression. A cross-validation R^2 can be defined by

$$R_{cv}^2 = 1 - \frac{PRESS}{SSTotal}$$

Unlike R^2, this R_{cv}^2 does not increase if irrelevant predictors are added to the model; rather it tends to decrease. And where R^2 is necessarily non-negative, R_{cv}^2 may be negative. This non-uncommon situation is an indication that the model fitted is poor – worse, in fact, than making predictions by ignoring the predictors and using the mean activity as the prediction in all circumstances.

R_{cv}^2 mimics the results of applying the final regression to predicting a future case; large values can be interpreted unequivocally and without regard to either the number of cases or predictors as indicating that the fitted regression will accurately predict the activity of future compounds of the same chemical type as those used to calibrate the regression.

3. RESULTS AND DISCUSSION

Table III provides results of studies done on the complete set of 31 diverse compounds as well as the subset of 18 haloalkanes for the prediction of human $\log P_{blood:air}$. Examining the models developed using structural descriptors, we find that the RR methodology is generally superior to both PCR and PLS. This is supported by our earlier studies with various congeneric and diverse sets of chemicals.(34-36) The model developed using TC descriptors as independent variables was superior to those developed with other structural descriptor classes in the analysis of the 31 diverse compounds, while the TS+TC model was superior in the analysis of the 18 haloalkanes.

The results of QSPRs reported in this paper show that structure-property correlations are comparable or superior to property-property correlations involving experimental saline:air and olive oil:air partition coefficients in the prediction of human blood:air partition coefficient. For the set of 31 diverse chemicals, a cross-validated R^2 of 0.874 and a PRESS of 7.79 is obtained for the TC model, while the property-property model utilizing $\log P_{saline:air}$ and $\log P_{olive:oil\ air}$ yields a cross-validated R^2 of 0.889 with a PRESS of 6.19 (Table III). For the set of 18 haloalkanes, the TS+TC models yields a cross-validated R^2 of 0.897 with a PRESS of 3.02, while the property-property model utilizing $\log P_{saline:air}$ and $\log P_{olive:oil\ air}$ yields a cross-validated R^2 of 0.846 with a PRESS of 4.50. However, property-property models in which rat $\log P_{blood:air}$ is used to predict human $\log P_{blood:air}$ are superior to those in which either $\log P_{saline:air}$ and $\log P_{olive:oil\ air}$ or structural parameters are used as predictors; with a cross-validated R^2 of 0.963 and PRESS of 2.25 for the full set of 31 compounds, and a cross-validated R^2 of 0.961 and PRESS of 1.16 for the subset of 18 haloalkanes.

It is clear from the results presented in Table III that experimental rat blood:air partition coefficient is the best predictor of human blood:air partition coefficient. Acquiring these data, however, is time consuming and requires laboratory testing resources along with the sacrifice of animals. Experimental determination of rat blood:air partition coefficient of hundreds or thousands of candidate chemicals would be a daunting task. The theoretical descriptor-based models, on the other hand, can provide reasonable estimates very quickly and at a low cost.

Ridge regression coefficients and standard errors for the top 10 descriptors based on | t | values for the human $\log P_{blood:air}$ TC model based on the set of 31 diverse chemicals are provided in Table IV. The indices most important for the prediction of human $\log P_{blood:air}$ include: a) molecular weight (fw), quantifying molecular size, b) triplet indices (AZV_y), encoding information about the nature of atoms, c) electrotopological state indices (SdO, SddSN, SSBr), which are numerical descriptors of the electronic states of atoms, d) valence and bonding connectivity indices (${}^1\chi^b$, ${}^1\chi^v$), which quantify structural information regarding molecular size and shape, and e) a hydrogen bonding parameter (HB_1). The important role of molecular factors such as size, electronic interactions, and hydrogen bonding in determining partition coefficients of chemicals is evident from our earlier studies[3, 37] and those of Kamlet et al.[38]

It is important to reiterate that model predictability is best judged, not with a fitted model, but with a cross-validated model wherein each of the compounds, in turn, is omitted from the data set and its value then determined by the coefficients of the remaining n-1 compounds. In this way, we have an accurate, if not conservative, indication of how well the model will predict property values of new compounds which are similar to those used to create the model. Figure 1 illustrates the relationship between the fitted and experimental human $\log P_{blood:air}$ values using the TC model for the set of 31 diverse compounds. All statistical values reported in this paper, however, are based on cross-validated results. Accordingly, Figure 2 illustrates the relationship between the cross-validated predicted and experimental human $\log P_{blood:air}$ values using the TC model for the set of 31 diverse compounds.

In conclusion, the models based on rat $\log P_{blood:air}$ are superior to any of the structure-based models. It is important to note, however, that experimental data are not currently available for the majority of compounds; and obtaining this data is costly in terms of time and monetary resources. In contrast, we are able to obtain reasonably good models using structural descriptors that can be calculated very quickly and inexpensively for both existing and unsynthesized chemicals. Modeling based on structural descriptors also promotes an understanding of the theoretical basis of properties and reduces the need for animal research, an area to which a growing aversion exists in our society.

ACKNOWLEDGEMENT

This is contribution number XXX from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported in part by Grant F49620-01-0098 from the United States Air Force.

REFERENCES

1. Cash, G. G. (2001). Personal communication.
2. Poulin, P. & Krishnan, K. (1996). A mechanistic algorithm for predicting blood:air partition coefficients of organic chemicals with the consideration of reversible binding in hemoglobin. *Toxicol. Appl. Pharmacol.*, 136, 131-137.
3. Niemi, G. J., Basak, S. C., Veith, G. D. & Grunwald, G. (1992). Prediction of octanol-water partition coefficient (Kow) using algorithmically-derived variables. *Environ. Toxicol. Chem.*, 11, 891-898.
4. Katritzky, A. R., Wang, Y., Sild, S. & Tamm, T. (1998). QSPR studies on vapor pressure, aqueous solubility, and the prediction of water-air partition coefficients. *J. Chem. Inf. Comput. Sci.*, 38, 720-725.
5. Basak, S. C., Gute, B. D. & Drewes, L. R. (1996). Predicting blood-brain transport of drugs: A computational approach. *Pharm. Res.*, 13, 775-778.
6. Gute, B. D., Grunwald, G. D. & Basak, S. C. (1999). Prediction of the dermal penetration of polycyclic aromatic hydrocarbons (PAHs): A hierarchical QSAR approach. *SAR QSAR Environ. Res.*, 10, 1-15.
7. Gargas, M. L., Burgess, R. J., Voisard, D. E., Cason, G. H. & Andersen, M. E. (1989). Partition coefficients of low molecular weight volatile chemicals in various tissues and liquids. *Toxicol. Appl. Pharmacol.*, 98, 87-99.
8. Sato, A. & Nakajima, T. (1979). Partition coefficients of some aromatic hydrocarbons and ketones in water, blood and oil. *Br. J. Ind. Med.*, 36, 231-234.
9. Parham, M., Hall, L. H. & Kier, L. B. (2000). LogP. www.logP.com.
10. Basak, S. C., Harriss, D. K. & Magnuson, V. R. (1988). POLLY, Version 2.3, Copyright of the University of Minnesota.
11. Filip, P. A., Balaban, T. S. & Balaban, A. T. (1987). A new approach for devising local graph invariants: Derived topological indices with low degeneracy and good correlational ability. *J. Math. Chem.*, 1, 61-83.

12. Basak, S. C., Balaban, A. T., Grunwald, G. D. & Gute, B. D. (2000). Topological indices: Their nature and mutual relatedness. *J. Chem. Inf. Comput. Sci.*, 40, 891-898.
13. Hall Associates Consulting, Molconn-Z Version 3.50, Quincy, MA, 2000.
14. Kier, L. B. & Hall, L. H. (1986). *Molecular Connectivity in Structure-Activity Analysis*, Research Studies Press, Letchworth, Hertfordshire, U.K.
15. Kier, L. B., Murray, W. J., Randić, M. & Hall, L. H. (1976). Molecular connectivity. V. Connectivity series concept applied to diversity. *J. Pharm. Sci.*, 65, 1226-1230.
16. Randić, M. (1975). On characterization of molecular branching. *J. Am. Chem. Soc.*, 97, 6609-6615.
17. Basak, S. C., Magnuson, V. R., Niemi, G. J. & Regal, R. R. (1988). Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math.*, 19, 17-44.
18. Raychaudhury, C., Ray, S. K., Ghosh, J. J., Roy, A. B. & Basak, S. C. (1984). Discrimination of isomeric structures using information theoretic topological indices. *J. Comput. Chem.*, 5, 581-588.
19. Basak, S. C. (1999). Information theoretic indices of neighborhood complexity and their applications. In *Topological Indices and Related Descriptors in QSAR and QSPR* (Devillers, J. and Balaban, A.T., Eds.) pp. 563-593, Gordon and Breach Science Publishers, The Netherlands.
20. Kier, L. B. & Hall, L. H. (1999). *Molecular Structure Description: The Electrotopological State*, Academic Press, San Diego, CA.
21. Hall, L. H., Mohney, B. & Kier, L. B. (1991). The electrotopological state: Structure information at the atomic level for molecular graphs. *J. Chem. Inf. Comput. Sci.*, 31, 76-82.
22. Basak, S. C. (1988). H-Bond, Copyright of the University of Minnesota.
23. Balaban, A. T. (1982). Highly discriminating distance-based topological indices. *Chem. Phys. Lett.*, 89, 399-404.
24. Balaban, A. T. (1983). Topological indices based on topological distances in molecular graphs. *Pure and Appl. Chem.*, 55, 199-206.
25. Balaban, A. T. (1985). Chemical graphs. Part 48. Topological index J for heteroatom-containing molecules taking into account periodicities of element properties. *Math. Chem. (MATCH)*, 21, 115-122.
26. SAS Institute, Inc. In SAS/STAT User Guide, Release 6.03 Edition; Cary, NC, 1988.
27. Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 8, 27-51.
28. Massy, W. F. (1965). Principal components regression in exploratory statistical research. *J. Am. Statistical Assoc.*, 60, 234-246.
29. Hoskuldsson, A. (1988). PLS regression methods. *J. Chemometrics*, 2, 211-228.
30. Hoskuldsson, A. (1995). A combined theory for PCA and PLS. *J. Chemometrics*, 9, 91-123.
31. Wold, H. (1975). Soft modeling by latent variables: The nonlinear iterative partial least squares approach. In *Perspectives in Probability and Statistics, Papers in Honor of M. S. Bartlett* (Gani, J., Ed.). Academic Press, London.
32. Miller, A. J. (1990). *Subset selection in regression*, Chapman and Hall, New York.
33. Rencher, A. C. & Pun, F. C. (1980). Inflation of R2 in best subset regression. *Technometrics*, 22, 49-53.
34. Hawkins, D., Basak, S. & Shi, X. (2001). QSAR with few compounds and many features. *J. Chem. Inf. Comput. Sci.*, 41, 663-670.
35. Basak, S. C., Hawkins, D. M. & Mills, D. (2002). Predicting blood:air partition coefficient of structurally diverse chemicals using theoretical molecular descriptors. In *Advances in Molecular Similarity*; Girones, X., Carbo-Dorca, R., Mezey, P. G., Eds.; Kluwer, in press.
36. Basak, S. C., El-Masri, H., Hawkins, D. M. & Mills, D. (2001). Exposure assessment of volatile organic chemicals(VOCs): Predicting blood:air partition coefficients of diverse chemicals using theoretical descriptors. *J. Chem. Inf. Comput. Sci.*, submitted.
37. Basak, S. C., Niemi, G. J. & Veith, G. D. (1990). Recent developments in the characterization of chemical structure using graph-theoretic indices. In *Computational Chemical Graph Theory* (Rouvray, D.H., Ed.). pp. 235-277.
38. Kamlet, M. J., Abboud, J.-L. M., Abraham, M. H. & Taft, R. W. (1983). Linear solvation energy relationships. 23. A comprehensive collection of the solvatochromatic parameters, π^*, α and β, and some methods for simplifying the general solvatochromatic equation. *J. Org. Chem.*, 48, 2877-2887.

Table I. Experimental liquid:air partition coefficients [a] and calculated $\log P_{n\text{-octanol:water}}$

		Experimental				Calculated
				Rat	Human	
No.	Chemical	P(0.9%saline:air)	P(olive oil:air)	P(blood:air)	P(blood:air)	LogP (*n*-octanol:water)
Haloalkanes						
1	Dichloromethane	5.96 ± 0.71	131 ± 7	19.4 ± 0.8	8.94 ± 0.13	1.16
2	Chloroform	3.38 ± 0.09	402 ± 12	20.8 ± 0.1	6.85 ± 0.51	1.86
3	Carbon tetrachloride	0.35 ± 0.03	374 ± 11	4.52 ± 0.35	2.73 ± 0.23	3
4	Chlorodibromomethane	7.34 ± 0.42	2683 ± 152	116 ± 4	52.7 ± 1.2	1.77
5	Chloroethane	1.09 ± 0.06	38.9 ± 3.1	4.08 ± 0.39	2.69 ± 0.20	1.47
6	1,1-Dichloroethane	2.45 ± 0.04	186 ± 7	11.2 ± 0.1	4.94 ± 0.24	1.86
7	1,2-Dichloroethane	11.4 ± 0.1	366 ± 8	30.4 ± 1.2	19.5 ± 0.7	1.6
8	1,1,1-Trichloroethane	0.75 ± 0.07	295 ± 22	5.76 ± 0.50	2.53 ± 0.13	2.26
9	1,1,2-Trichloroethane	13.3 ± 0.3	1776 ± 26	58.0 ± 1.1	35.7 ± 0.4	2.08
10	1,1,1,2-Tetrachloroethane	3.53 ± 0.23	2686 ± 51	41.7 ± 1.0	30.2 ± 1.3	2.64
11	1,1,2,2-Tetrachloroethane	23.4 ± 2.0	6358 ± 402	142 ± 6	116 ± 6	2.51
12	Hexachloroethane	0.66 ± 0.21	5015 ± 318	62.7 ± 2.1	52.4 ± 1.4	4.24
13	1-Bromo-2-chloroethane	8.91 ± 0.56	569 ± 23	52.7 ± 3.5	29.2 ± 2.1	1.73
14	1-Chloropropane	1.04 ± 0.01	105 ± 2	5.21 ± 0.06	2.85 ± 0.06	1.95
15	2-Chloropropane	0.82 ± 0.09	69.9 ± 3.5	3.10 ± 0.17	1.39 ± 0.29	1.81
16	1,2-Dichloropropane	2.75 ± 0.11	428 ± 30	18.7 ± 0.5	8.75 ± 0.50	2.18
17	*n*-Propyl bromide	1.44 ± 0.12	272 ± 8	11.7 ± 0.4	7.08 ± 0.40	2.13
18	Isopropyl bromide	1.08 ± 0.04	164 ± 5	5.95 ± 0.14	2.57 ± 0.15	1.63
19	1-Nitropropane	127 ± 4	1062 ± 21	223 ± 10	187 ± 6	0.8
20	2-Nitropropane	98.3 ± 5.4	640 ± 16	183 ± 12	154 ± 17	0.61
21	*n*-Heptane	0.18 ± 0.10	405 ± 3	4.75 ± 0.15	8.19 ± 0.10	4.31
22	JP-10 (tricyclo[5.2.1.0 ^{2,6}]-decane)	0.21 ± 0.07	12970 ± 420	62 ± 4	52.5 ± 3.7	3.75
23	Vinyl chloride	0.43 ± 0.04	24.4 ± 3.7	1.68 ± 0.18	1.16 ± 0.08	1.37
24	Trichloroethylene	0.83 ± 0.30	553 ± 46	21.9 ± 1.4	8.11 ± 0.17	2.36
25	Tetrachloroethylene	0.79 ± 0.06	2134 ± 159	18.9 ±1.1	10.3 ± 1.1	3.47
26	Vinyl bromide	0.44 ± 0.06	56.0 ± 1.5	4.05 ± 0.16	2.27 ± 0.16	1.61
27	Benzene	2.75 ± 0.10	465 ± 5	17.8 ± 0.3	8.19 ± 0.10	2.04
28	Chlorobenzene	2.81 ± 0.07	2188 ± 41	59.4 ± 1.0	30.0 ± 0.3	2.64
29	*o*-Xylene	2.65 ± 0.08	3534 ± 208	44.3 ± 2.0	34.9 ± 1.7	3.15
30	*m*-Xylene	1.92 ± 0.12	3245 ± 116	46.0 ± 1.5	32.5 ± 1.6	3.21
31	*p*-Xylene	1.77 ± 0.07	3319 ± 96	41.3 ± 3.5	44.7 ± 1.9	3.20

[a] Values represent mean ± standard error

Table II. Symbols, definitions and classification of calculated molecular descriptors

	Topostructural (TS)
I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\bar{I}_D^W$	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
$\overline{IC}$	Information content of the distance matrix partitioned by frequency of occurrences of distance h
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
${}^h\chi$	Path connectivity index of order h = 0-10
${}^h\chi_C$	Cluster connectivity index of order h = 3-6
${}^h\chi_{PC}$	Path-cluster connectivity index of order h = 4-6
${}^h\chi_{Ch}$	Chain connectivity index of order h = 3-10
P_h	Number of paths of length h = 0-10
J	Balaban's J index based on topological distance
nrings	Number of rings in a graph
ncirc	Number of circuits in a graph
DN^2S_y	Triplet index from distance matrix, square of graph order (# of non-H atoms), and distance sum; operation y = 1-5
DN^21_y	Triplet index from distance matrix, square of graph order, and number 1; operation y = 1-5
$AS1_y$	Triplet index from adjacency matrix, distance sum, and number 1; operation y = 1-5
$DS1_y$	Triplet index from distance matrix, distance sum, and number 1; operation y = 1-5
ASN_y	Triplet index from adjacency matrix, distance sum, and graph order; operation y = 1-5
DSN_y	Triplet index from distance matrix, distance sum, and graph order; operation y = 1-5
DN^2N_y	Triplet index from distance matrix, square of graph order, and graph order; operation y = 1-5
ANS_y	Triplet index from adjacency matrix, graph order, and distance sum; operation y = 1-5
$AN1_y$	Triplet index from adjacency matrix, graph order, and number 1; operation y = 1-5
ANN_y	Triplet index from adjacency matrix, graph order, and graph order again; operation y = 1-5
ASV_y	Triplet index from adjacency matrix, distance sum, and vertex degree; operation y = 1-5
DSV_y	Triplet index from distance matrix, distance sum, and vertex degree; operation y = 1-5
ANV_y	Triplet index from adjacency matrix, graph order, and vertex degree; operation y = 1-5
	Topochemical (TC)
O	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
O_{orb}	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-suppressed graph
I_{orb}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
IC_r	Mean information content or complexity of a graph based on the rth (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for rth (r = 0-6) order neighborhood of vertices in a hydrogen-

	filled graph
CIC_r	Complementary information content for rth (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
${}^{h}\chi^{b}$	Bond path connectivity index of order h = 0-6
${}^{h}\chi_{C}^{b}$	Bond cluster connectivity index of order h = 3-6
${}^{h}\chi_{Ch}^{b}$	Bond chain connectivity index of order h = 3- 6
${}^{h}\chi_{PC}^{b}$	Bond path-cluster connectivity index of order h = 4-6
${}^{h}\chi^{v}$	Valence path connectivity index of order h = 0-10
${}^{h}\chi_{C}^{v}$	Valence cluster connectivity index of order h = 3-6
${}^{h}\chi_{Ch}^{v}$	Valence chain connectivity index of order h = 3-10
${}^{h}\chi_{PC}^{v}$	Valence path-cluster connectivity index of order h = 4-6
J^B	Balaban's J index based on bond types
J^X	Balaban's J index based on relative electronegativities
J^Y	Balaban's J index based on relative covalent radii
HB_1	Hydrogen bonding parameter
AZV_y	Triplet index from adjacency matrix, atomic number, and vertex degree; operation y = 1-5
AZS_y	Triplet index from adjacency matrix, atomic number, and distance sum; operation y = 1-5
ASZ_y	Triplet index from adjacency matrix, distance sum, and atomic number; operation y = 1-5
AZN_y	Triplet index from adjacency matrix, atomic number, and graph order; operation y = 1-5
ANZ_y	Triplet index from adjacency matrix, graph order, and atomic number; operation y = 1-5
DSZ_y	Triplet index from distance matrix, distance sum, and atomic number; operation y = 1-5
DN^2Z_y	Triplet index from distance matrix, square of graph order, and atomic number; operation y = 1-5
nvx	Number of non-hydrogen atoms in a molecule
nelem	Number of elements in a molecule
fw	Molecular weight
si	Shannon information index
totop	Total Topological Index t
sumI	Sum of the intrinsic state values I
sumdelI	Sum of delta-I values
tets2	Total topological state index based on electrotopological state indices
phia	Flexibility index (kp1* kp2/nvx)
IdCbar	Bonchev-Trinajsti□ information index
IdC	Bonchev-Trinajsti□ information index
Wp	Wienerp
Pf	Plattf
Wt	Total Wiener number
knotp	Difference of chi-cluster-3 and path/cluster-4
knotpv	Valence difference of chi-cluster-3 and path/cluster-4
nclass	Number of classes of topologically (symmetry) equivalent graph vertices
numHBd	Number of hydrogen bond donors
numwHBd	Number of weak hydrogen bond donors
numHBa	Number of hydrogen bond acceptors
SHCsats	E-State of C sp^3 bonded to other saturated C atoms
SHCsatu	E-State of C sp^3 bonded to unsaturated C atoms
SHvin	E-State of C atoms in the vinyl group, =CH-
SHtvin	E-State of C atoms in the terminal vinyl group, =CH_2
SHavin	E-State of C atoms in the vinyl group, =CH-, bonded to an aromatic C
SHarom	E-State of C sp^2 which are part of an aromatic system
SHHBd	Hydrogen bond donor index, sum of Hydrogen E-State values for –OH, =NH, -NH2, -NH-, -SH, and #CH
SHwHBd	Weak hydrogen bond donor index, sum of C-H Hydrogen E-State values for hydrogen atoms on a C to which a F and/or Cl are also bonded
SHHBa	Hydrogen bond acceptor index, sum of the E-State values for –OH, =NH,

	-NH2, -NH-, >N-, -O-, -S-, along with –F and –Cl
Qv	General Polarity descriptor
NHBint$_y$	Count of potential internal hydrogen bonders (y = 2-10)
SHBint$_y$	E-State descriptors of potential internal hydrogen bond strength (y =2-10)
	Electrotopological State index values for atoms types: SHsOH, SHdNH, SHsSH, SHsNH2, SHssNH, SHtCH, SHother, SHCHnX, Hmax Gmax, Hmin, Gmin, Hmaxpos, Hminneg, SsLi, SssBe, Sssss,Bem, SssBH, SsssB, SssssBm, SsCH3, SdCH2, SssCH2, StCH, SdsCH, SaaCH, SsssCH, SddC,StsC, SdssC, SaasC, SaaaC, SssssC, SsNH3p, SsNH2, SssNH2p, SdNH, SssNH, SaaNH, StN, SsssNHp, SdsN, SaaN, SsssN, SddsN, SaasN, SssssNp, SsOH, SdO, SssO, SaaO, SsF, SsSiH3, SssSiH2, SsssSiH, SssssSi, SsPH2, SssPH, SsssP, SdsssP, SsssssP, SsSH, SdS, SssS, SaaS, SdssS, SddssS, SssssssS, SsCl, SsGeH3, SssGeH2, SsssGeH, SssssGe, SsAsH2, SssAsH, SsssAs, SdsssAs, SsssssAs, SsSeH, SdSe, SssSe, SaaSe, SdssSe, SddssSe, SsBr, SsSnH3, SssSnH2, SsssSnH, SssssSn, SsI, SsPbH3, SssPbH2, SsssPbH, SssssPb
	Geometrical / Shape (3D)
kp0	Kappa zero
kp1-kp3	Kappa simple indices
ka1-ka3	Kappa alpha indices

Table III. Summary statistics of predictive models for human $\log P_{\mathrm{blood:air}}$ based on experimental properties and theoretical structural descriptors.

A. 31 DIVERSE CHEMICALS

Independent Variables	RR		PCR		PLS		LR	
	$R^2_{\mathrm{c.v.}}$	PRESS	$R^2_{\mathrm{c.v.}}$	PRESS	$R^2_{\mathrm{c.v.}}$	PRESS	$R^2_{\mathrm{c.v.}}$	PRESS
Structural descriptors								
TS	0.257	45.8	-0.451	89.4	0.052	58.4		
TS+TC	0.846	9.48	0.165	51.4	0.677	19.9		
TS+TC+3D	0.827	10.6	0.140	53.0	0.620	23.4		
TS+TC+3D+logP[a]	0.835	10.2	0.112	54.7	0.652	21.4		
TS	0.257	45.8	-0.451	89.4	0.052	58.4		
TC	0.874	7.79	0.403	36.8	0.709	17.9		
3D	0.147	52.6	-0.013	62.4	-0.256	77.4		
Properties								
$\mathrm{LogP}_{\mathrm{olive\ oil:air}} + \mathrm{LogP}_{\mathrm{saline:air}}$							0.899	6.19
Rat $\log P_{\mathrm{blood:air}}$							0.963	2.25

B. 18 HALOALKANES

Independent Variables	RR		PCR		PLS		LR	
	$R^2_{\mathrm{c.v.}}$	PRESS	$R^2_{\mathrm{c.v.}}$	PRESS	$R^2_{\mathrm{c.v.}}$	PRESS	$R^2_{\mathrm{c.v.}}$	PRESS
Structural descriptors								
TS	0.252	22.0	-1.53	74.3	-0.815	53.2		
TS+TC	0.897	3.02	0.825	5.14	0.678	9.45		
TS+TC+3D	0.892	3.16	0.856	4.22	0.702	8.74		
TS+TC+3D+logP[a]	0.892	3.18	0.856	4.23	0.704	8.69		
TS	0.252	22.0	-1.53	74.3	-0.815	53.2		
TC	0.891	3.21	0.853	4.32	0.616	11.3		
3D	0.753	7.24	0.593	11.9	0.562	12.9		
Properties								
$\mathrm{LogP}_{\mathrm{olive\ oil:air}} + \mathrm{LogP}_{\mathrm{saline:air}}$							0.846	4.50
Rat $\log P_{\mathrm{blood:air}}$							0.961	1.16

[a] Calculated $\log P_{n\text{-octanol:water}}$; values included in Table I.

Table IV. Ridge regression coefficient and standard error for each of the top 10 descriptors, ranked by | t |, in the topochemical model for the prediction of human log$P_{blood:air}$, n = 31.

Descriptor	RR coeff	s.e.	t
SdO	0.227	0.021	10.690
HB_1	0.340	0.032	10.660
SddsN	-1.694	0.159	-10.640
AZV_3	0.130	0.016	8.000
${}^1\chi^v$	0.345	0.052	6.670
AZV_4	0.224	0.034	6.580
AZV_1	0.133	0.024	5.640
SsBr	0.238	0.044	5.390
fw	0.287	0.054	5.310
${}^1\chi^b$	0.139	0.028	5.060

FIGURE CAPTIONS

Figure 1. Experimental *vs* fitted human $\log P_{blood:air}$ using the topochemical (TC) ridge regression (RR) model for the set of 31 diverse compounds

Figure 2. Experimental *vs* cross-validated predicted human $\log P_{blood:air}$ using the topochemical (TC) ridge regression (RR) model for the set of 31 diverse compounds

Figure 1.

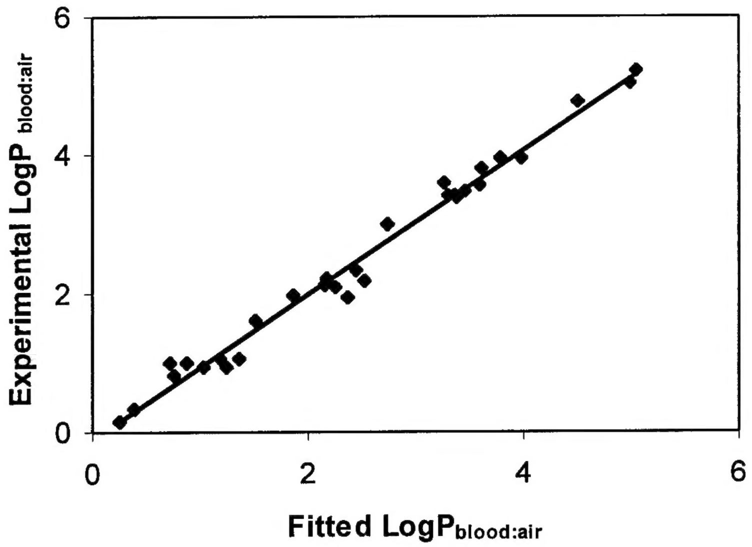
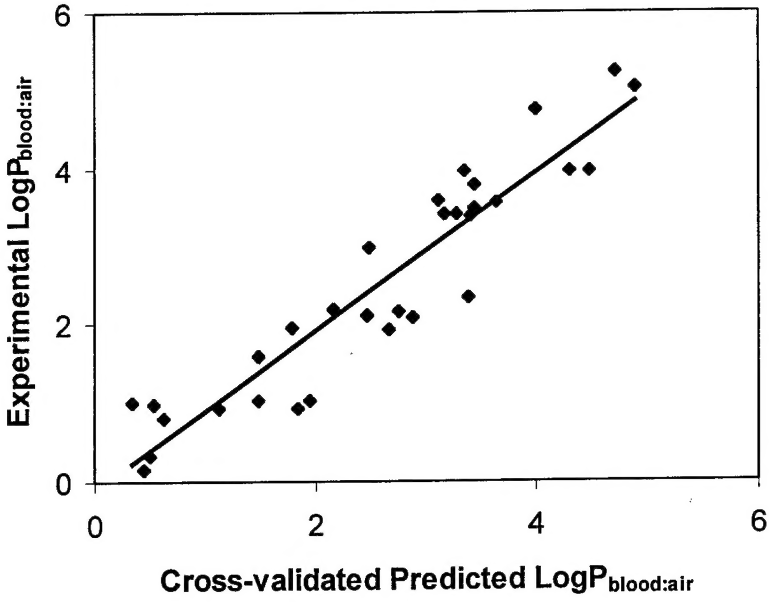

Experimental LogP blood:air
Fitted LogPblood:air
0
2
4
6
0
2
4
6

Figure 2.



6
4
2
0
0
2
4
6
Experimental LogP$_{blood:air}$
Cross-validated Predicted LogP$_{blood:air}$